DOCUMENT RESUME

ED 230 578 TM 830 283

AUTHOR Gamache, LeAnn M.
TITLE Comparison of Traditional and Latent Trait Procedures in Analysis and Selection of Rating Scale Items.
PUB DATE 14 Apr 83
NOTE 68p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Comparative Analysis; Correlation; Higher Education; *Item Analysis; Item Banks; *Latent Trait Theory; *Rating Scales; Scores; Test Construction; *Test Reliability; *Test Theory; *Test Validity
IDENTIFIERS Birnbaum Models; Dichotomous Variables; Guttman Scales; *Item Rating Scale; Likert Scales; Rasch Model; Thurstone Scales

ABSTRACT

Scales constructed under procedures and criteria outlined by the various traditional and latent trait methods were examined as to whether they varied in characteristics related to scale quality. Scales were constructed from a common pool of items analyzed in full form according to Likert and a one-parameter Rasch model for non-dichotomous data. Scales were analyzed in dichotomous form for the Guttman, two-parameter Birnbaum, and one-parameter Rasch models. Findings indicate that, for this data set, a traditional method based on item to total score correlation produced a slightly more valid scale. All five method-defined scales, however, were remarkably similar in other characteristics related to scale quality. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED230578

COMPARISON OF TRADITIONAL AND LATENT TRAIT PROCEDURES IN ANALYSIS AND SELECTION OF RATING SCALE ITEMS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
LeAnn M. Gamache
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

LeAnn M. Gamache
The University of Iowa

A paper presented at the Annual Meeting of the National Council on Measurement in Education, April 14, 1983, Montreal, Canada.

TM 830 283

ABSTRACT

The major purpose of the study was to examine whether scales constructed under procedures and criteria outlined by the various traditional and latent trait methods varied in characteristics related to scale quality. Scales were constructed from a common pool of items analyzed in full form according to Likert and a one-parameter Rasch model for non-dichotomous data and analyzed in dichotomous form for the Guttman, two-parameter Birnbaum, and one-parameter Rasch models. Findings indicated that, for this data set, a traditional method based on item to total score correlation produced a slightly more valid scale. All five method-defined scales, however, were remarkably similar in other characteristics related to scale quality.

INTRODUCTION

Measurement in the area of aptitude and achievement deals with the most directly observable of the human behaviors and is relatively sophisticated. Most investigation, however, is related to noncognitive characteristics and involves the use of questionnaires, judges' ratings, self-report ratings, interviews, and similar procedures. Dawes (1972) reports that 60% of the studies conducted involve dependent variables measured solely with a form of rating procedure.

Traditional approaches to scale construction and analysis include Thurstone's equal appearing interval approach, the summated rating technique used with Likert-type items, and Guttman cumulative scale procedures. Briefly, the Thurstone approach entails the ranking of statements by judges as to degree of trait assessed and subsequent assignment of a numerical value representing the location of the item on the trait continuum. The scale is then administered to subjects who check statements with which they agree (Mehrens & Lehmann, 1975, p. 379). A person's score is the mean or median of the values of the statements endorsed. The scale is independent of the sample with which it is used; similarly Thurstone indicates that individuals' scores are free of the particular items used (Andrich, 1980, p.11).

Likert-type items require the examinee to indicate degree of intensity for each statement. Much less difficult to construct, this type of instrument is most frequently utilized. A person's score is the simple sum of integer values given for each response. Likert rejected the Thurstone idea that the properties of the instrument be unaffected by characteristics of the sample of examinees and that the response process follow a formal statistical model (Andrich, 1980).

The Guttman cumulative scale technique is a procedure developed to evaluate sets of statements against the criteria for an underlying unidimensional scale (Edwurds, 1957, p. 172), an assumption underpinning all three traditional approaches to scale development (Mehrens & Lehmann, 1975, p. 381). The task for the examinee is identical to that of the Likert technique, but an attempt is made to order both examinees and items with respect to the trait continuum. Examinees and items are thus assigned scale values indicating position along the underlying continuum (Torgerson, 1958, p.47). From each person's rank, the procedure checks the assumption that the pattern of response to each item can be reproduced precisely. Each deviation from the expected pattern is considered error, which if sufficiently large when accumulated across all examinees, refutes the unidimensionality and cumulative nature required of a Guttman scale

(Edwards, 1957; Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

Various issues remain unresolved by these traditional procedures with regard to the qualities requisite for ideal measurement. With the Likert approach there is concern that scores are not directly interpretable and must be examined in light of a criterion group. The absence of a zero point handicaps the direct use of individual scores (Edwards, 1957, p.157). The adequacy of integer scoring rests upon the validity of the assumption of equal distances between response categories and is not often verified (Andrich, 1980; Liebermann, 1971). The disposition of a middle "undecided" category has also caused some confusion regarding appropriate scoring due to concerns of possible response set by examinees (Andrich, 1980). The reservations regarding the Thurstone and Guttman procedures stem basically from the complexities of the respective procedures. In addition, studies by Hovland & Sherif (1952) and Thurstone & Chave (1929), indicate scale values assigned in the Thurstone procedure are significantly related to judges' attitude or level of the trait being calibrated (Mehrens & Lehmann, 1975, p.380). Guttman procedures rarely seem to verify scales of more than a half dozen items (Torgerson, 1958).

A host of problems inherent to ratings, particularly self-report ratings, compound concerns related to the spe-

cific scale development procedures. Instruments employing self-report rating scale techniques are vulnerable to a variety of invalidating factors. Instruments carefully developed along textbook guidelines (Mehrens & Lehmann, 1975, p. 378; Payne, 1974; Thorndike & Hagen, 1977, p. 436) as well as published inventories resulting from years of development activity are affected by problems of definition of the construct assessed, examinee response sets, faking, and the unwillingness or inability of the examinee to accurately self-report. These inherent problems result in measurements of lower reliability and validity than is considered acceptable for cognitive assessment instruments. It is for these reasons that noncognitive self-report measurement is subject to considerable criticism (Mehrens & Lehmann, 1975, p.537).

Many of the techniques and criteria applied to development of rating scale instruments emerged from the more structured situation for cognitive measurement. Perhaps recent advances in cognitive measurement techniques with use of latent trait analysis may also be generalizable to development of measurement instruments for noncognitive characteristics to better exhibit the characteristics of ideal measurement. Latent trait models specify a mathematical relationship between observable examinee test performance and the unobservable, latent trait assumed to structure

examinee performance (Hambleton & Cook, 1977). The mathematical function designates the probability of correct response by an examinee of a given ability level to a particular item. Logistic latent trait models are of most interest in the literature and are the only type considered here. The number of parameters required to estimate this function distinguishes three major latent trait models. Birnbaum's three-parameter model considers item difficulty, item discrimination, and examinee guessing in determining ability along the latent trait continuum. The two-parameter model is concerned with item difficulty and discrimination. The one-parameter Rasch model is the simplest of the latent trait models, considering item difficulty to be the only parameter necessary to determine examinee position along the trait continuum. These models make various assumptions concerning the test data. It is generally agreed that different models are appropriate for different sets of data, and one should find the model that fits the data, although Wright (1968) argues that tests should be designed to fit the assumptions of the model to be used. All models assume the underlying trait is unidimensional, that only one unobservable characteristic structures examinee response to items. This is equivalent to a requirement of local independence of each item. The two-parameter model in addition assumes guessing has not influenced examinee response to

items. The Rasch model further assumes homogeneity of item discrimination (Gustafsson, 1977; Hambleton, 1977), absence of guessing, and that the examinees had sufficient opportunity to respond to each item.

Forms of latent trait analysis have been applied to rating scale or ordinal data in studies by Liebermann (1971), Schmitt (1981), and Andrich (1978). Liebermann scaled nine stories assessing moral development along Kolberg's six stage continuum. The probability that an examinee at a given level of development would answer with a pure pattern of responses provided category scores for each item. More recently, Schmitt analyzed the three-alternative Central Life Interest scale by dichotomizing the continuum and applying the Rasch one-parameter model for dichotmous items. Andrich developed a model for latent trait analysis of rating scale data then selected sixteen statements from a set of twenty based on the results of the analysis for inclusion in a final version of a questionnaire. Criteria for inclusion required the thresholds between alternatives be correctly ordered on the item characteristic curve and that the distances between categories within the item be symmetrical.

Various models have been presented extending latent trait analysis to the nondichotomous situation. In addition to the work of Bock & Jones (1968), Strivastava & Webster (1967), Samejima (1969), and Kalakowski & Bock (1972) as

reported by Liebermann (1971) and Andrich (1980), Masters (1981) has extended a Rasch model derived by Andrich to the situation where category alternatives are free to vary in number and structure from item to item. The CREDIT computer program (Masters, Wright, & Ludlow, 1980) analyzes rating scale or other ordinal alternative data in terms of the steps or distances along the continuum between the alternatives. The model assumes that each successive step "taken" along the continuum of the alternatives represents a greater degree of the variable being measured. The level of the individual on the trait continuum is independent of the position or difficulty of the items used, as with the dichotomously scored items analyzed with the BICAL program (Wright & Mead, 1978). This member of the family of Rasch models is claimed to share the unique potential that item parameters are estimated independently of the calibration sample and that both measurement of persons and analysis of items is freed from the specific set of items and persons used for calibration (Masters, 1981).

What is the practical impact on scale quality, if any, of employing these varying methods of scale analysis and construction? The major purpose of the study was to examine whether scales constructed under procedures and criteria outlined by the various traditional and latent trait methods varied in characteristics related to scale quality. Scales

were constructed from a common pool of items analyzed in full form according to Likert and the one-parameter Rasch model for nondichotomous data and analyzed in dichotomous form for the Guttman, two-parameter Birnbaum, and one-parameter Rasch models.

REVIEW OF LITERATURE

METHODS OF ITEM ANALYSIS AND SELECTION

Likert Analysis

With a Likert analysis, subjects alone receive scores. Items are selected so as to increase individual differences with respect to the attitude or trait continuum (Torgerson, 1958, p.47). Different weights might be given to items on the basis of their power to discriminate between examinees, not on the basis of their relative position along an underlying continuum. Items with the highest item to total scale correlations are selected to best differentiate among examinees.

Guttman Analysis

Guttman analysis is appropriate for sets of items defined a priori to be of homogeneous content and of a unidimensional nature. Although item analysis is not a necessary part of Guttman analysis, sets of statements found to be internally consistent (Edwards & Kilpatrick, 1948) are more apt than unselected items to meet the Guttman require-

ments that only one pattern of responses corresponds to a given total score and that thus the scale is cumulative.

The problem is not one of testing whether a perfect scale exists, since one variant observation is sufficient to reject such a claim. Instead, the purpose of the analysis is to determine whether the set of items is sufficiently unidimensional to enable treatment of the data as if it were a perfect scale (Torgerson, 1958, p. 322-323). The procedure ranks items as to proportion of examinees selecting each category. It is expected that examinees with the highest total scores constitute the group of examinees endorsing the least frequently endorsed items, i.e. items assessing a high level of the trait of interest. It is further expected that if an examinee endorses an item indicating a high degree of the trait, he will also ensdorse items less highly ranked along the trait continuum. The degree to which such a pattern of responses can be determined from the total score is the degree of reproducibility of responses. Cut points are set for each item to mark the location where the most common response shifts from nonendorsement to endorsement. Two guidelines apply to the selection of these cut points: (1) they are to be located so as to minimize error, defined as an unexpected endorsement or nonendorsement response to the category, and (2) no category should have more error than nonerror (Edwards, 1957).

The percent accuracy of predicting or reproducing item scores from total scores is reflected in the coefficient of reproducibility. Each deviation of the observed item response from that expected from the total score is considered an error (e). These errors are summed across examinees and items then divided by the product of the number of items (k) and examinees (N) to determine the proportion of error (E). The difference between the proportion of error (E) and one constitutes the coefficient of reproducibilty (R) for the total scale (Edwards, 1957, p.195).

$$R = 1 - (\Sigma\Sigma e/Nk) \qquad (1)$$

Individual item coefficients of reproducibility are similarly calculated by summing across examinees and dividing by the number of examinees. Goodenough (1944) and Edwards (1948) have improved the coefficient of reproducibility to take into account response patterns for other items as well as the total score (Edwards, 1957, p. 184).

The question of whether the error of prediction is random or reflects the presence of a second variable underlying response to the set of items is then investigated. The actual patterns of response of the examinees is screened for frequent but unexpected patterns. The presence of such nonscale patterns in the response of examinees indicates the existance of at least one other variable underlying the

scale (Guttman, 1947). By definition a Guttman scale has only one pattern of response for each total score. There are 1+(nk)-n desirable response patterns for a set of n items having k response categories for each item.

A coefficient of reproducibility of .90 or greater (Edwards, 1957, p. 185) is necessary but not sufficient evidence of the scalability of a set of statements. Because the reproduciblity can never be less than the frequency of the modal response, the expected value of the coefficient of reproducibility is high especially when the number of response categories is small. It is further affected by the number of items in the set (Torgerson, 1958, p. 322-323) Furthermore, the coefficient is spuriously high to the degree the responses are unevenly distributed across the response categories. The minimal marginal reproducibility (M), therefore, is determined from the proportion (P) of responses in the modal category for each statement (Edwards, 1957, p. 193).

$$M = \Sigma P/k \qquad (2)$$

The difference between the coefficient of reproducibility (R) and the minimum marginal reproducibility (M) is termed the percent improvement and indicates the extent to which the coefficient of reproducibility is due to the response distribution rather than the inherent cumulative relation-

ship of items. Finally, the coefficient of scalability (S) is calculated as the percent improvement divided by the largest possible improvement.

$$S = (R - M)/(1 - M) \qquad (3)$$

The coefficient of scalability should be greater than .60 if the scale is truly unidimensional and cumulative (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975, p. 533).

Several criteria are outlined to assure the coefficient of reproducibility is not spuriously high (Torgerson, 1958, p. 324).

1. The more response categories within an item, the more confidence can be placed in the coefficient of reproducibility. If dichotomous items are used, at least ten items should be included in the scale; fewer multicategory items would suffice.
2. Although it is desirable to have a considerable range of item marginals to provide a range of scores, extreme marginals inflate the coefficient of reproducibility. Few if any items should exhibit more than 80% of the responses in the modal category. If inclusion of such an item in the scale is deemed necessary, it should be excluded in calculation of the coefficient of reproducibility.

3. The pattern of errors in prediction should be random. A large number of examinees should not be found to exhibit the same nonscale pattern of responses. Existance of such patterns indicates more than one variable underlies the scale.
4. Individual items should have item reproducibilities greater than .85.
5. Finally, each item category should have more nonerror than error.

If these criteria are met, the statement content is scalable and the properties of a perfect scale apply. If all criteria except that pertaining to the coefficient of reproducibility are met, the set of items can be termed a quasi-scale and judged to reflect only unsystematic error. Scores on such a scale are assumed to be determined by one major

Latent Trait Analysis - Birnbaum Model

Birnbaum's three-parameter latent trait model (Birnbaum, 1968) encompasses both the Rasch and the two-parameter latent trait models. The three-parameter model indicates an examinee's probability of scoring on item g to be a function of the lower asymptote of the item characteristic curve or chance level (c_g), the trait level of the examinee (θ), and the level of the item (b_g), and the item's power to discriminate (a_g):

$$P_g(\theta) = c_g + (1-c_g)\frac{\exp(Da_g(\theta-b_g))}{1+(Da_g(\theta-b_g))} \qquad (4)$$

The guessing parameter (c_g) is the asymptotic probability of scoring on the item for examinees with a trait level at some negligible level. Although this parameter may be higher or lower than the usual notion of chance as the reciprocal of the number of item alternatives, it is bounded by 0 and 1. The item level parameter (b_g) is defined as the point on the latent continuum scale at which examinees located at the trait level have a probability of $(1+c_g)/2$ of scoring on the item. The discrimination parameter (a_g) is proportional to the slope of the item characteristic curve at the point where examine trait level is equal to item level. It is calculated as $Da_g(1-c_g)/2$ where D is a constant. Setting D to equal 1.7 enables the logistic function to most closely match the normal distribution function (Kolen, 1979).

Conditional estimation procedures are used to maximize the likelihood function of responses of N examinees to n items. The process of alternatively fixing examinee trait level estimates and item parameter estimates is repeated until a given standard of stability is attained. The estimation procedure utilized by LOGIST (Wood, Wingersky, & Lord, 1978) includes four steps, each reaching convergence defined as an increase in the criterion likelihood function

of less than the default value of .2 percent or some other user specified standard. Due to insufficient information, estimates of trait level cannot be made for examinees with perfect scores, scores of zero, and scores of less than 33%. Such examinees are deleted from the calibration process. Best calibration is obtained when more than forty items and one thousand examinees are used (Wood & Lord, 1976).

Item quality is judged in terms of the amount of information contributed by an item regarding the trait level of examinees (Lord, 1980). For a given trait level (θ) and dichotomous item score (u_i) of 0 or 1, the information function is:

$$I(\theta,u_i) = \frac{P'_i{}^2}{P_iQ_i} \qquad (5)$$

where P_i is the probability $P(\theta)$ of a scored response at a given trait level, Q_i is the complement of that probabilty or the probability of a scoreless response, and P'_i is the derivative of P_i with respect to theta (θ) (Lord, 1980, p. 21). Item information then is given by:

$$I(\theta,u_i) = D^2a_i^2 \frac{Q_i}{P_i}\left[\frac{P_i-c_i}{1-c_i}\right]^2 \qquad (6)$$

where D is a constant, usually 1.7, a_i is the item's discrimination parameter, and c_i is the pseudo-chance level or

guessing parameter for the item (Lord, 1980, p. 73). The item information function can also be obtained from:

$$I(\theta,u_i) = \frac{D^2a_i^2(1-c_i)}{(c_i+e^{DL_i})(1+e^{-DL_i})^2} \qquad (7)$$

where $Li=a_i(\theta-b_i)$ (Lord, 1980, p. 73).

The test information function is the sum of the item information functions for all items included in the test. When items are optimally weighted, the contribution of an item to the measurement effectiveness of the total test does not depend on characteristics of other items included in the test. This contrasts with the classical test theory approach where the contribution of an item to reliability or validity is dependent upon other items in the test (Lord, 1980, p. 22). The optimal scoring method for maximum information is the linear composite of weighted dichotomous scores with weights of P_i'/P_iQ_i (Lord, 1980, p.23). When the two-parameter model is employed, this optimal scoring weight is independent of examinee level (θ). In such situations, the optimal weight becomes the item's discrimination (Lord, 1980, p. 77).

To construct a test, the ideal shape of the test information function must be specified in advance for a target group of examinees. Items are selected so their item information curves fill the area under the desired test information for the target group. Items are added and/or replaced

until a satisfactory approximation of the desired test information curve is obtained (Lord, 1980, p. 23).

Latent Trait Analysis - Rasch Model for Dicotomous Data

The model suggests that an examinee's response to an item is governed only by the examinee's level of the trait and the trait level of the item. Thus the Rasch model, indicating probability of earning a score on an item of given trait level by examinee of a given level, is the logistic function

$$P(x_{vi}=1|\beta_v\delta_i) = \exp(\beta_v-\delta_i)/(1+\exp(\beta_v-\delta_i)) \qquad (8)$$

where β_v represents the examinee's trait level and δ_i represents the level of the item. Notation used here conforms to that of the literature regarding the Rasch model, although the model is equivalent to the Birnbaum model when the guessing parameter is set to zero and the discrimination parameter is held constant. If both the examinee and item are at the same level along the underlying trait continuum, the probability of the examinee scoring on the item is .5. The greater the difference between the examinee's position along the trait continuum and the trait level of the item, the greater the departure from .5 of the probability of earning a score on the item.

Four guidelines for examining item quality are applied to information in the BICAL (Wright & Mead, 1978) output and summarized in the total t-fit statistic: (1) Item characteristic curve information gives the proportion scoring at each of six groups of examinee trait levels. The proportion is expected to increase with each higher trait level. (2) There should be no differences in magnitude or direction between trait level groups in departure from expected response. (3) Tests of fit should indicate invariance of item level estimates across trait level groups. (4) The discrimination index should not depart substantially from the model expectation of 1.0. These four guidelines for judging item appropriateness are summarized in the t-fit statistic (Wright & Stone, 1977):

$$t = [\ln(v) + v - 1][f/8]^{\frac{1}{2}} \qquad (9)$$

where $v=\Sigma z^2/f$ and f is one less than the number of examinees for item fit and represents one less than the number of items when used for person fit analysis. Item misfit is indicated with an absolute value greater than 3.0 for the t-fit statistic. Application of the specific criteria can be made to diagnose the cause of misfit.

Items with an absolute t-fit value of less than 3.0 form a calibrated pool from which to select items most appropriate for a specific measurement situation. The cali-

bration of item level allows selection of items to best match a presumed trait level, range, and distribution for a target group of examinees. The objective of item selection is to combine such items so as to minimize the standard error of measurement (Wright & Stone, 1979, p. 131).

When items operate as expected by the Rasch model, only consideration of item level parameters is necessary in determining the operating characteristics of the resulting test. Three test parameters are required to characterize a test: the average level of the selected items, the range of item levels, and the number of items included in the test. Birnbaum's analysis of information maximization implies that a test intended for a normally distributed group of examinees be composed of normally distributed item levels. A uniform test, in which item levels are evenly spaced from least to greatest degree of trait (Wright & Stone, 1979, p. 466), is more analogous to physical measurement with a yardstick. Wright & Douglas (1975a) found the normal test best for normal target group of examinees, and the interval test to work best for a target group for which the trait distribution was unknown. When comparing the two types of tests for a normal distribution target group, measurement precision was deemed equivalent for all practical purposes and the uniform test considered the best all purpose test dstribution of item levels. For such a distribution, a set of items is desired to have item levels (D_i) such that

$$Di=H-(W/2)[(L-2i+1)/L] \quad (10)$$

for items i= 1 to L where H represents the hypothesized location of the group of examinees along the trait continuum, W is the desired width of the test approximated as four times the hypothesized standard deviation of trait levels within the group, and L is the number of items to be included calculated as the ratio of the error coefficient function of the given width and expected score proportion to the square of the desired the standard error of measurement (Wright & Stone, 1979, p. 140). Calibrated items are selected from the pool that most closely match the set of item levels deemed ideal for the target group.

Latent Trait Analysis - Rasch Model for Polychotomous Data

The Partial Credit Model (Masters, 1981) for multicategory response data is another member of the family of Rasch models. With some expansion and redefinition, this model is identical to the Rasch model for dichotomously scored items and used in a similar manner for analysis of nondichotomous items.

The CREDIT computer program (Masters, Wright, & Ludlow, 1980) is based on the Partial Credit Model and parallels BICAL in general functioning. Persons with perfect scores

and scores of zero are deleted from calibration as are items on which all persons scored in the highest or lowest category. Fit statistics are calculated in the same manner for persons and items. An absolute value of the t-fit statistic greater than 3.0 is the criterion for misfit. Such misfitting persons and items have not responded to the measurement situation as did other persons and items. Unlike BICAL, however, CREDIT does not exclude misfitting persons from the calibration process. From the calibrated pool of items meeting the fit criteria, items are selected to best match the set of trait levels calculated as best matching the characteristics of the target group, again using the formula $Di=H-(W/2)[(L-2i+1)/L]$.

SOME COMPARISONS OF PROCEDURES

Comparisons have been made of the procedures and characteristics of test resulting from application of these various methods for item analysis and selection. Andrich (1980) indicates that the Rasch model for nondichotomous data serves to unify the traditional Likert and Thurstone procedures, retaining the theoretical and practical features of both methods. As with the Thurstone procedure, statement scale values are determined independently of the trait level of examinees, probability statements can be made regarding fit of the data to the model, and scale values need not be

equal interval in nature. As with the Likert procedure, the total score is a sufficient statistic for locating an examinee's level of trait along the latent continuum, and time consuming a priori judgments as to item level are not required. Although infrequently applied, the main features of Guttman analysis could also be incorporated through further development of the model (Andrich, 1980).

Ryan and Hamm (1976) contrasted Rasch analysis to procedures prescribed by classical test theory which suggest that the internal consistency form of reliability is increased by lengthening a test with parallel items. Rasch analysis identifies multidimensional items, although results will not replicate those of a factor analysis. If the identified items are deleted, the shortened test has higher reliability than the original test. They conclude such a procedure provides a method of increasing the reliability of a test after it has been administered to examinees.

Bejar (1977) examined the question of whether a latent trait approach to test construction with items of a continuous response nature improved reliability of the resultant scale over that resulting from test construction according to a classical approach with dichotomous items. Bejar applied the continuous response model to sets of personality items then mathematically argued that the continuous response was 57% more informative, as defined by Lord, than the

maximally informative dichotomous item for these data. The benefit increased with the difference between examinee trait level and the level at which the item yielded maximum information. This would imply that scales composed of items selected from analysis and scoring of items in nondichotomous form would yield validity coefficients greater than those for scales resulting from analysis, scoring, and selection of items in dichotomous form. He concluded that the extension of latent trait theories to polychotomous and continuous response situations is of practical importance since these conditions allow more precise estimation of person position along the trait continuum.

Benson, Crocker, and Ware (1978) compared classical item analysis, factor analysis, and Rasch analysis techniques for development of a verbal aptitude subtest in terms of internal consistency, standard error, item difficulty, item discrimination, and test efficiency characteristics of the resulting tests. No differences in standard error or reliability of the resulting forms were found. The relative efficiency, as defined by Lord (1980), of the tests, however, varied by examinee ability level. Factor analysis yielded greater efficiency than classical analysis for the lowest ten percentile scorers. Rasch analysis, with selection of items having the best mean square fit values, produced a test of greater efficiency than classical

analysis for those in the 40-90%ile range, but less efficiency for the extreme scorers. The test developed from Rasch analysis had greater efficiency than that from the factor analysis for the 50-90%ile group; factor analysis yielded greater efficiency for the 1-20%ile and 98-99%ile groups of examinees.

METHOD

The purpose of the study was to examine whether scales constructed under the procedures and criteria outlined by the various traditional and latent trait approaches varied in characteristics related to scale quality. Scales were constructed according to the various procedures and compared with respect to score equivalence, scale reliability, and scale validity.

The data consisted of the responses of 505 University of Iowa students to the UCLA Revised Loneliness Scale as part of a questionnaire submitted to them in the spring of 1980 requesting information related to demographics, social relationships, coping behaviors, and self-esteem (Russell, Stephan, & Salih, 1981). These respondents represented a 50% return of 200 students randomly selected at each of the freshman, sophomore, junior, senior, and graduate levels. Respondents were approximately evenly distributed across level in school with somewhat fewer seniors and graduate

students responding. Ninety-five percent of the respondents was white. Fifty-five percent of the sample was female. These variables and other demographic characteristics of the sample are given in Table 1 with available data concerning the population sampled.

The Revised UCLA Loneliness Scale (Russell, Peplau, & Cutrona, 1980; Russell, Peplau, & Ferguson, 1978) consists of twenty items related to the subjective experience of loneliness, which were selected from a larger pool based on item to total score correlations. Responses related to frequency of feelings such as "no one knows me well" and "there are people I can talk to" are made to a four point scale. The direction of the response categories is reversed for ten of the twenty items.

The first phase of the study analyzed the responses of a random sample of 404 respondents using each of five methods for item analysis and selection in scale development: a summated rating procedure accomplished by the LIKERT and RESPONSE COUNTS scoring programs available at the University of Iowa Evaluation and Examination Service; a Guttman scalogram analysis of the dichotomized responses by the GUTTMAN SCALE program from the SPSS statistical package (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975); latent trait analysis of dichotomous data using LOGIST for Birnbaum's two-parameter model (Wood, Wingersky, & Lord, 1976); latent trait ana-

lysis of the dichotomized data using BICAL for a Rasch one-parameter model (Wright & Mead, 1978); and a Rasch one-parameter latent trait analysis for multiple category data using the CREDIT program (Masters, Wright, & Ludlow, 1980).

The GUTTMAN program is limited to the analysis of twelve items. The scale, therefore, was separated into two sets containing items 1 through 12 and 9 through 20 respectively for analysis. Items 9 through 12 were included in both scalings to provide information as to comparability of the underlying scale for the sets of items (Torgerson, 1958). The response continuum was dichotomized at 2.5 for the GUTTMAN analysis. Use of the LOGIST and BICAL computer programs for latent trait analyses also required that the item responses be dichotomized. Responses to categories 1 and 2 were combined to form one category as were responses to categories 3 and 4. The third parameter, "c", was set to 0 since guessing is not a factor influencing response to this type of item. Program defaults were taken for each method of analysis. The criterion for misfit of both item and persons in the BICAL and CREDIT analyses was set at an absolute value of a t-fit statistic greater than 3.0 as suggested by Wright and Stone (1979).

Independent analyses of the results of each method provided a description of the characteristics of the Loneliness Scale items and diagnostic information for item selection.

Five shortened versions of the scale were constructed with the ten items best meeting the statistical guidelines for item selection of the respective methods. First, items not meeting any dichotomous criteria were discarded. If fewer than ten items remained, the least offending items were returned to the pool. The remaining items were ranked with respect to each of the other criteria and assigned an average rank across all nondichotomous criteria. The ten items most highly ranked were identified for inclusion in the shortened scale representing the analysis procedure. The item selection procedure for the Likert analysis, therefore, was to rank items according to their item to total scale correlations and select the top ten for inclusion in the LIKERT scale. For the GUTTMAN scale, items having more than 80% of the examinees responding in the modal category were deleted. The remaining items were ranked according to proportion error from least to greatest and according to item coefficient of reproducibility (Eq. 1) from highest to lowest. The ten items having the highest average of these rankings were included in the scale. Item selection for the LOGIST scale consisted of deleting items for which no estimate of trait level or discrimination could be made, then determining the information (Eq. 7) provided by each item across trait levels, estimated to range from positive to negative 3.0 for the target population. Items were selected

such that when information was accumulated across items the test information curve would be a straight line, indicating equal information was desired at each trait level. The procedure for selection of items for the BICAL scale consisted of first deleting items having absolute t-fit statistic values greater than 3.0. From the remaining items, items were selected that had trait levels nearest to the set determined as ideal for the target group of examinees (Eq. 10). The target group was assumed to be identical to the calibration group with respect to trait level and dispersion. Selection of items for the CREDIT scale paralleled that used for BICAL. Items with an absolute t-fit value greater than 3.0 were deleted. An ideal set of item trait levels was determined (Eq. 10) under the assumption the target group had trait level and dispersion characteristics identical to the calibration group. The ten items having trait levels closest to the ideal trait levels were then selected.

In the second phase of the study, raw scores for the 101 respondents reserved from the calibration sample were determined for each of the five resultant versions of the scale. Analysis was restricted to the descriptive level due to overlap of items in the five scales. Concurrence of item selection and item rankings, correlations between examinee scores across scales, coefficient alpha reliabilities, and

correlations with criterion were examined for the five scales. The loneliness criterion used was a total score, LINDEX, of seven items external to the Revised UCLA Loneliness Scale, which directly asked respondents to indicate their loneliness characteristics along a nine category response continuum. Validity of the method-defined scales and other referent scales was determined in relation to the LINDEX score. In addition, reliability and validity coefficients and the correlation matrix of scores were examined with reference to several other scales. A random selection of ten of the twenty items was used to create the RANDOM scale. The ten items having the highest average ranks across all methods were combined as the BESTAVG scale; the remaining ten items created scale WORSTAVG. The ten items not selected within each method were identified as scales NLIKERT, NGUTTMAN, NLOGIST, NBICAL, and NCREDIT to provide unique referent scales for the LIKERT, GUTTMAN, LOGIST, BICAL, and CREDIT scales respectively.

RESULTS

ITEM ANALYSIS AND SELECTION

LIKERT Item Analysis and Selection

Examination of the LIKERT item analysis and RESPONSE COUNTS output provided information regarding item means, standard deviations, item to scale correlations, and item

response frequencies. Table 2 provides this information for the twenty items analyzed.

Item means on the four-point response continuum ranged from 1.35 to 2.49; standard deviations for these items ranged from .63 to 1.0; item to total scale correlations ranged from a low of .26 to .78. Most item responses were heavily weighted at the nonlonely end of the response continuum. Examinees essentially responded as if to a three category continuum. Only one item attracted more than ten percent of the respondents to the highest category. Six items exhibited 1% or less of the group responding in the loneliest response category. Four items had more than 50% of the respondents selecting the least lonely of the four categories.

The ranking of these items based upon item to scale correlation (Edwards, 1957, p. 155) resulted in the selection, in order of preference, of items 14, 18, 11, 13, 2, 7, 17, 19, 16, and 3 for the shortened version based on LIKERT analysis.

GUTTMAN Item Analysis and Selection

GUTTMAN analysis was performed on two subsets of items dichotomized at the midpoint of the response category continuum. The first twelve items and the last twelve items were separately scaled. Items 9 through 12 were included in

both analyses for comparison of the subsets of items. The results of these analyses are presented in Table 3.

The proportion of examinees endorsing the item represents the traditional concept of difficulty. Item endorsements ranged from 7% for Item 1 to 51% for Item 13. The proportion of error for each item represents the proportion of unexpected endorsements and nonendorsements across examinees given their respective scores on the total collection of items. Both the proportion of error and its complement, the item coefficient of reproducibility, depend upon the other items included in the analyses. The proportion of error ranged from .05 for Item 19 to .35 for Item 4. These items, therefore, exhibited the highest and lowest coefficients of reproducibility respectively. The percent of examinees in the modal category is also indicated.

Analysis of the first twelve items indicated a coefficient of reproducibility of .8082; minimum marginal reproducibility was .7581. The percent improvement, therefore, was .0501 and the coefficient of scalability was .2070. This subset of items did not have the characteristics of a GUTTMAN scale. The coefficient of reproducibility was due to the high degree of concurrence in item endorsement across examinees. Similarly, the second subset of twelve items did not meet the criteria for a GUTTMAN scale. The coefficient of reproducibility for this set was .8473 with a minimum

marginal reproducibility of .7432, percent improvement of .1040, and a coefficient of scalability of .4051. Criteria of a GUTTMAN scale require a coefficient of reproducibility greater than or equal to .90 and a coefficient of scalability greater than .6. Because the subsets could not meet the criteria of a GUTTMAN scale, the total set of twenty items could not exhibit GUTTMAN scale characteristics.

In examining the proportion of error and the point biserial correlations of the four items analyzed as part of both subsets of items, the performance of three of the four items appears to be quite similar. The average of the values from the two analyses was used for these four items in the ranking of items against the selection criteria.

Items were examined against the item criteria outlined for GUTTMAN scale items: (1) less than 80% of the responses accumulated in the modal response category; (2) items exhibited item reproducibilities of at least .85; and (3) response to each item category contained more nonerror than error, i.e. that the proportion of error for a dichotomous item was less than .5. Application of the absolute criterion of percent of respondents in the modal category resulted in the rejection of nine items as indicated by an "x" in Table 3. The average of the ranking of the relative criteria for the remaining items indicate the best ten items in order of desirability under this method are Items 14, 7,

17, 3, 11, 2, 18, 13, 8, and 12. Of all items in the pool only Item 14 met all criteria.

LOGIST Analysis and Item Selection

LOGIST analysis of the dichotomized items with reference to the two-parameter model was made on the 351 examinees remaining in the calibration sample after removal of those with perfect scores, scores of zero, and those examinees scoring on less than seven of the twenty items. Eighteen iterations were required to reach convergence, defined as an increase in the maximum likelihood function of less than .2 percent. The latent standard score scale, defined to have a mean of 0 and a standard deviation of 1, was anchored on the 335 examinees having trait levels within the range of positive to negative 3.0. The final item parameters are given in Table 4.

A stable estimate of the discrimination parameter of Item 4 was negative or could not be made. Under such circumstances, the program substitutes a value of .01 for the discrimination parameter for calibration purposes. Values for the discrimination parameter ranged from the low of less than .01 for Item 4 to a high of 1.56 for Item 14. Item trait levels on the standard score scale, defined by examinee levels, ranged from a low loneliness level of -0.42 for Item 13 to a high level of loneliness of 2.37 for Item 1, disregading the inestimable Item 4.

The level of information was calculated for trait levels from -3.0 to +3.0 at intervals of .4 scale units for each item. The total of these information values across the trait levels is also given by item in Table 4. The specific values for the information index for the selected examinee levels is provided in Table 5.

The plots of information by trait level for each item were drawn to determine the best combination of items to fill the area under the desired test information curve. It was assumed the target population would be similar to the calibration sample in trait level and dispersion. The desired test information curve for the shortened ten-item version of the test was to provide the highest possible degree of information across all trait levels ranging from -3 to +3 standard deviations and would thus be a straight line. The best approximation to the desired test curve included Items 14, 11, 18, 13, 2, 17, 3, 7, 12, and 8 in order of the amount of information contributed. The resulting test information caurve is compared to the desired curve in Figure 1. The desired test information curve could not be developed from the items in the pool. This set of items is identical to that selected from results and criteria applied in the GUTTMAN analysis.

BICAL Analysis and Item Selection

BICAL analysis of the twenty dichotomized items according to the Rasch model was based upon 349 of the 404 examinees remaining after the editing process and intial calibration deleted 52 examinees having total scores of zero, 2 examinees exhibiting perfect total scores, and one examinee evidencing an implausible response record with a misfit value greater than 3.0. No items were deleted from calibration due to insufficient information caused by lack of variation in examinee response. The 349 examinees in the calibration sample exhibited an average trait level of -1.41 logits with a standard deviation of 1.08 logits. Item information in Table 6 for each item includes item trait level, standard error of trait level estimates, a discrimination index, total t-fit statistic, and error impact. The discrimination index indicates the linear trend of departure from the model across trait level groups around the model value of 1.0. The total t-fit statistic indicates agreement between the variable defined by the item with that defined by all other items in the set. Misfit is defined as an absolute value of the t-fit statistic greater than 3.0. Error impact is the proportion of error increase in the test due to misfit of the item.

Item trait levels ranged from a low of -1.94 logits for Item 13 to a loneliness level of 2.16 logits for Item 19.

Item levels were accompanied by standard errors from .13 to .26 logits. Item 4 exhibited negative discrimination; only six items were within 10% of the expected value of 1.0. Item fit values ranged from -4.91 to 7.55 for Items 14 and 4 respectively. Items 4, 8, 9, and 12 were identified as contributing relatively more error for the collection of items calibrated.

Item quality guidelines that the proportion of examinees scoring on an item increase with trait level, that there be no difference in magnitude or direction of the departure of the observed from expected proportion across trait level subgroups and that the discrimination index not differ substantially from 1.0 are summarized in the t-fit statistic. Items 14 and 4 have absolute fit values greater than 3.0 are were discarded from the item pool.

Final selection of items for the shortened version was made according to the item trait level, the only item parameter assumed by the model to influence response of an examinee. Selection of these items was based on the assumption that the target population of examinees was like the calibration sample in trait level and dispersion. The set of items to match the target examinee group, therefore, was to satisfy the equation (Di) for items i = 1 to 10:

$$Di = -1.41 - (1.08/2)[(10-2i+1)/10]$$

The ideal set of items would be those having trait levels of -1.9, -1.79, -1.68, -1.57, -1.46, -1.36, -1.25, -1.14, -1.03, and -0.92 logits. The best approximation for this set that could be made from items remaining in the pool in order of preference were Items 13, 18, 12, 8, 2, 11, 17, 7, 3, and 16, having trait levels of -1.94, -1.54, -1.35, -1.27, -1.19, -1.15, -0.70, -0.01, 0.05, and 0.19 respectively.

CREDIT Analysis and Item Selection

Analysis of the responses to the twenty items using the CREDIT program for nondichotomous data was made on the 388 response records of examinees remaining in the calibration sample after the editing process for missing data, perfect scores, and scores of zero was completed. Fifteen records were deleted due to missing response to one or more items. One examinee record was deleted due to a total score of zero and thus insufficient information for estimation of his trait level. No items were excluded due to lack of variability in response. CREDIT does not exclude misfitting examinees from the calibration process as does BICAL.

Trait level estimates for the steps taken within an item were -.55, -.14, and .70 logits. These within-item step distances were assumed constant across all items. The item is anchored at the zero logit point of the within-item

continuum to the latent continuum for the entire set of items. The four category response continuum was not an equal interval scale according to this calibration; an increase of nearly double the degree of loneliness was required to take the step from category 3 to category 4 than to take the step from category 2 to category 3.

The distribution of the loneliness trait levels for examinees was quite skewed in the negative direction with a mean of -1.30 and a standard deviation of 1.13 logits. Scores ranged from a high of 2.38 logits to a low of -4.62 logits. The group of examinees exhibited a much lower degree of loneliness than that assessed by the items. The mean fit value for the 388 examinees was -.14 with a standard deviation of 1.49. Fourteen examinees exhibited absolute fit values greater than 3.0, indicating they did not respond to the measurement situation as did other examinees in the calibration sample.

Because the trait level scale is defined in terms of the item trait levels, the mean item level was 0 with a standard deviation of .83 logits. Item levels range from a low loneliness level of -1.22 logits for Item 13 to a high level of loneliness of 1.37 logits for Item 19 as shown in Table 7. The mean error in item level calibration was .08 logits. The mean value for fit of items to the latent scale was -0.42 with a standard deviation of 3.90. Seven items

misfit with absolute fit values greater than 3.0. Items 11, 14, 2, 18, 19, 17, and 4 were therefore discarded from the item pool.

The target group of examinees for the shortened version was assumed to be equivalent to the calibration sample in trait level and dispersion. The function used to determine the ideal set of item trait levels therefore was for items i=1 to 10:

$$Di=-1.30 - (1.13/2)[(10-2i+1)/10]$$

The desired set of item trait levels was -1.81, -1.70, -1.58, -1.47, -1.36, -1.24, -1.13, -1.02, -0.90, and -0.79. Only three items remaining in the pool possessed trait levels within this range. The best available set of items, in order of preference, was composed of Items 13, 12, 8, 3, 9, 16, 7, 6, 5, and 1. The trait levels for these items were -1.22, -0.97, -0.86, 0.03, 0.09, 0.15, 0.21, 0.43, 0.65, and 0.79 respectively.

COMPARISON OF RESULTANT TEN ITEM SCALES

The items selected for each method and ranks within the respective scales are reported in Table 8. There was considerable overlap in the items selected. Three items were selected by all five methods; five addtional items were represented in four of the five shortened scales. Items 4,

10, 15, and 20 were excluded from all scales. These four items all had the response continuum reversed to control possible response set of examinees and represented 40% of such items in the twenty item pool.

The GUTTMAN and LOGIST procedures produced scales containing the same items, all of which were worded to a positive response continuum. The LIKERT procedure produced a scale having 8 items in common with the GUTTMAN/LOGIST scale, 8 common items with BICAL and 4 items in common with the CREDIT scale. The GUTTMAN/LOGIST procedures produced a scale having an overlap of 9 items with BICAL and a 5 item overlap with CREDIT. The two Rasch procedures, BICAL and CREDIT, had six of the ten items in common.

The rankings of selected items within each method, however, varied across methods. The agreement between item rankings is given by Spearman rank order correlations given in Table 9.

These correlations more specifically reflect the agreement between the five procedures in item selection. Of some interest are the correlations of item rankings between LIKERT and CREDIT procedures which dealt with items in full form, between GUTTMAN, LOGIST and BICAL which analyzed items in dichotomous form, and between BICAL and CREDIT which followed the same latent trait model and differed only in the dichotomizing of the item response continuum.

Several other scales were constructed to serve as referent scales for examining the five scales. The mean and standard deviation for each of the four method-defined scales and the seven referent scales are given in Table 10. The items having the highest average rank and included in BESTAVG were the same items as selected by the GUTTMAN and LOGIST procedures. Scores for the 101 target examinees reserved from the calibration sample were calculated for each method-defined and referent scale using a sum of response category values as in the LIKERT procedure. In all comparisons, the means and standard deviations of the four method-defined scales were greater than those of the referent scales.

Coefficient alpha reliabilities are also given in Table 10. The original twenty items appear to have been rather homogeneous in that little variability was found in reliability coefficients across the various scales of selected and unselected items. The scale composed of randomly selected items had a reliability of .861. Reliability coefficents ranged from a low of .790 for the NLIKERT to a high of .868 for the LIKERT. Within each method, selected items generally exhibited slightly higher reliability than those unselected. With the CREDIT analysis, however, the selected items produced a scale having slightly lower reliability than its unselected counterpart.

A condensed matrix of the correlation between scores on the method-defined and referent scales is given in Table 11. It was considered unnecessary to correct these correlations for attenuation due to the degree of similarity in the respective reliabilities. All correlations between the method-defined scales would be 1.0 when corrected for attenuation. Correlations between the method-defined scales and the randomly selected set of items are also 1.0 when corrected for attenuation of the reliabilities. Examining the correlations between the unique sets of selected and unselected items within each method indicates the effect of applying the respective technique for item selection. The greatest difference in scores for scales so defined is found with the GUTTMAN/LOGIST procedures for this set of items. The only conclusion to be drawn, however, is that the item selection procedures do distinguish sets of items.

The validity coefficients for the method-defined and referent scales, calculated as correlation with total LINDEX score, are provided in Table 12, again with the reliability coefficients.

Validity coefficients for the shortened scales ranged from .800 for LIKERT to .671 for CREDIT. Coefficients for the referent scales ranged from .797 for NCREDIT to .638 for NLIKERT. The validity coefficients must be examined in light of desired levels of reliability, since the question

is whether the criterion measure can appropriately be replaced by the measures under investigation. The LIKERT procedure produced scales of selected and unselected items exhibiting the highest and lowest validities respectively. The random selection of items resulted in the next most valid scale. Four of the five procedures produced scales with validities greater than the sets of items not selected by the procedures. The CREDIT procedure, however, resulted in defining a scale of considerably less validity than items not selected for inclusion.

DISCUSSION

Scales resulting from the five item analysis and selection procedures were remarkably similar in characteristics indicative of scale quality. This was due to the considerable overlap of items selected under each procedure. Scores across the various scales were equivalent for the target group of examinees. Little variation was found in the internal reliability estimates between the shortened scale versions. Estimates of validity defined as the correlation with the external LINDEX score, however, reflected some effect of method of item selection. The LIKERT method provided the most valid scale, although only slightly more valid than that resulting from the random selection of items. The scales resulting from latent trait analysis of

dichotomized items evidenced equal validity when corrected for the reliability of the scales. The scale resulting from analysis and selection according to the latent trait analysis of items in nondichotomous form had the lowest validity of the method-defined scales.

A random selection of items from the pool resulted in a scale having reliability and validity similar to or exceeding that of the method defined scales. From the comparison of selected and unselected items within each procedure it can be seen that, with the exception of the CREDIT procedure, some benefit was gained from application of the procedure for purposes of item analysis and selection. The superiority of a particular procedure, however, was not demonstrated.

The impact of the methods for item analysis and selection was limited by the relative homogeneity of items and size of the item pool, leading to extensive overlap in the sets of ten items selected by each method. Another limitation might be argued due to the summated rating scoring of response of the validation sample to the method-defined scales. Use of this LIKERT scoring in analyzing the resultant shortened scales might be said to bias the analysis in favor of the LIKERT procedure. In the hypothetical situation of ongoing use of these shortened scales, however, users would generally rely on a summated scoring procedure

due to constraints of sample size, expertise, and scoring resources. Nevertheless, the power of the latent trait methods is diluted when restricting the study in this way. The potential benefits of eliminating aberrant examinees from the data set, equating of forms, application of alternate conceptions of error and stability, and determination of examinee scores along the latent continuum could not be realized in this study.

Some details unrelated to the major thrust of the study are worthy of note:

1. The scales resulting from GUTTMAN and LOGIST analyses, in addition to being identical in terms of items selected, were composed strictly of positively worded items. All of the positive items from the original pool were selected by these procedures; all negatively worded items were excluded. Some characteristic of the subset of items led to differential response to the positively worded items than to the negatively worded items. This result is in accord with recent findings (Schriesheim & Hill, 1981) of significant differences in response accuracy between conditions of positively and negatively worded questionnaire items. That study concluded that although inclusion of negatively worded items may theoretically offset response set tendencies of examinees,

the actual effect may be to impair response accuracy and reduce response validity.

2. The CREDIT procedure typically found the most discriminating items to misfit and discarded them from consideration in the final item selection. The basic rationale was that such items are too idiosyncratic in relationship to characteristics of the calibration sample to be useful in measurement of the trait in the general population. The same is true of the BICAL procedure at a less obvious level due to the dichotomizing of the examinee responses for analysis. To challenge this rationale, scales resulting from application of these five procedures should be applied to a sample differing greatly from the calibration sample and again examined with respect to agreement of scores, scale reliability, and validity.
3. Results of the CREDIT item analysis indicated the within-item response continuum differed substantially from the equal interval scale assumed when the total score is the sum of response category values. Under such circumstances total scores of equal numerical value may not represent the same degree of trait. An examination of the improvement in validity and reliability characteristics of the scales under conditions of response weighting would be of interest,

particularly in light of preliminary analyses indicating that the within-item step distances varied greatly between items in contradiction to the assumption made by the CREDIT analysis of rating scale data. Examination might be made as to whether measurement precision and validity is improved by applying some weighting procedure based upon the step values unique to each item.

Results of this study both contradict and confirm earlier research findings and theory related to comparison of methods. First, the superior reliability of Rasch defined scales over the scale produced from LIKERT analysis, which was expected from the reports of Andrich (1980) and Ryan and Hamm (1976), was not substantiated. Second, the study did not confirm the improved validity of the CREDIT analysis as was expected from reports (Bejar, 1977) of increased information from latent trait analysis of continuous and polychotomous response items over that of dichotomous items. Some confirmation was found of the findings of Benson et al (1978) that a test produced from Rasch analysis was surpassed by classical analysis for examinees located in the lower half of the distribution of scores. The severe negative skew of the distribution of loneliness scores of examinees in this study and the superior validity of the scale developed with the LIKERT procedure over those developed from the Rasch procedures support their finding.

Any future investigation into the comparative benefits of applying these procedures in the development and use of rating scale instruments might profit from scoring responses according to the respective models. The initial item pool should be sufficiently large that only a very small proportion of the pool is to be selected for each scale. Further, a procedure based on Samejima's Graded Response Model (Samejima, 1969) extending the three-parameter latent trait model to the nondichotomous situation should be included among the procedures examined. The benefit of using the two-parameter latent trait model for item analysis and selection could then be examined without the loss of information resulting from dichotomizing the response continuum necessary for application of LOGIST.

The question remains as to the practical impact on scale quality of applying the various methods of item analysis and selection for development of rating scale instruments. Unless future research provides an indication that application of a specific model and method results in significantly improved reliability and validity, the choice of procedure for item analysis and selection will continue to be determined by the developer's greater familiarity with or interest in a given procedure and the availablity of the necessary technological resources.

Table 1

Demographic Characteristics of
Respondents and Population Sampled

DEMOGRAPHIC CHARACTERISTICS	PERCENT OF SAMPLE	PERCENT OF POPULATION*
SEX		
Female	55.4	46.1
Male	44.6	53.9
YEAR IN SCHOOL		
Freshman	22.2	17.2
Sophmore	22.0	14.4
Junior	21.0	17.1
Senior	18.0	14.3
Graduate	16.9	33.7
ETHNIC BACKGROUND		
Asian	2.0	0.7
Black	1.2	2.4
Caucasian	95.2	95.3
American Indian	0.8	0.4
Other	0.8	1.2
TYPE OF HOUSING		
Dormitory	44.0	24.5
Fraternity or sorority	6.7	5.4
Parents' home	3.2	2.1
Married student housing	3.0	4.1
Apartment, house, other	46.1	52.9
DISTANCE TO CAMPUS		
On campus	43.0	28.6
Within city limits	45.9	52.9
Outside city limits	10.1	11.3
Other	0.0	6.2
ROMANTIC STATUS		
Single	85.0	81.0
Married	15.0	19.0

*From "A Profile of Students Enrolled at The University of Iowa, First Semester 1979-80", Office of the Registrar, The University of Iowa.

Table 2

Item Analysis Information
for LIKERT Analysis

Item	Correl to Total	Mean	S.D.	*	Frequencies 1	2	3	4
1	.53	1.52	.63	*	.53	.38	.06	.00
2	.67	2.31	.81	*	.15	.44	.33	.06
3	.65	1.83	.90	*	.44	.33	.16	.06
4	.26	1.91	.89	*	.39	.34	.21	.05
5	.56	1.58	.78	*	.58	.27	.11	.02
6	.64	1.68	.75	*	.48	.37	.13	.01
7	.66	1.78	.91	*	.51	.26	.17	.05
8	.53	2.29	.91	*	.22	.36	.31	.09
9	.50	1.80	.80	*	.40	.42	.13	.04
10	.60	1.38	.65	*	.69	.22	.07	.01
11	.70	2.31	.77	*	.14	.47	.33	.05
12	.58	2.34	.84	*	.16	.41	.34	.08
13	.68	2.49	.98	*	.19	.28	.35	.16
14	.78	2.20	.87	*	.22	.42	.27	.07
15	.58	1.49	.69	*	.60	.30	.07	.01
16	.65	1.79	.83	*	.43	.36	.16	.03
17	.66	2.03	1.00	*	.38	.28	.22	.09
18	.72	2.35	.90	*	.19	.35	.35	.09
19	.66	1.35	.58	*	.69	.25	.05	.00
20	.63	1.41	.67	*	.68	.22	.08	.01

Table 3

GUTTMAN Analysis of Item Characteristics

Item	Average Rank	Percent Endorsmt	Propor. Error	Coef of Repro.	Percent in Modal
1	x	7	.0639	.936	93
2	6	40	.2122	.788	60
3	4	21	.1816	.818	79
4	11	26	.3529	.647	74
5	x	14	.1253	.875	86
6	x	14	.1099	.890	86
7	2	22	.1560	.844	78
8	9	41	.2967	.703	59
9	x	17	.1501*	.850*	83
10	x	8	.0675*	.933*	92
11	5	39	.1832*	.817*	61
12	10	42	.3324*	.668*	58
13	8	51	.2684	.732	51
14	1	35	.1443	.856	65
15	x	9	.0684	.932	91
16	x	19	.1519	.848	81
17	3	32	.1772	.823	68
18	7	45	.2430	.757	55
19	x	5	.0456	.954	95
20	x	9	.0709	.929	91

*Average value from both subscalings

Table 4

Item Parameters and
Information from Logist
Analysis

Item	Discrim.	Trait Level	Information Total
.1	.78	2.37	2.47
2	.99	0.10	4.10
3	.95	1.01	3.92
4	.01*	50.89	.00
5	.72	1.77	2.59
6	1.15	1.33	4.76
7	.93	.99	3.77
8	.39	.14	1.28
9	.47	2.03	1.43
10	.93	2.11	3.34
11	1.10	0.13	4.65
12	.50	0.03	1.87
13	1.02	-0.42	4.29
14	1.56	0.26	6.60
15	.83	2.13	2.89
16	.73	1.29	2.81
17	.94	0.47	3.93
18	1.06	-0.14	4.50
19	1.03	2.36	3.57
20	1.00	1.88	3.84

*Value substituted by LOGIST for negative discrimination

Table 5

Information Index at Selected Trait Levels (in Hundredths)

Item	Negative Trait Levels								Positive Trait Levels						
	3.0	*	*	*	*	*	*	*	*	*	*	*	*	*	3.0
1				1	1	2	3	5	9	14	21	30	38	43	43	37
2	2	3	6	11	19	33	50	64	68	58	41	26	15	8	4	2
3		1	1	3	5	9	17	28	44	59	65	59	45	29	17	10
4																
5		1	1	2	3	5	7	11	17	23	30	36	37	35	29	22
6				1	2	4	8	17	34	60	86	95	78	50	27	14
7		1	2	3	5	10	17	29	43	57	62	56	43	28	17	10
8	4	5	6	7	9	10	10	11	11	11	10	9	8	7	6	5
9	1	2	2	3	4	5	6	8	10	12	14	15	16	16	15	14
10				1	1	2	3	6	11	19	31	46	59	62	54	39
11	1	2	4	9	18	34	57	80	87	72	48	27	14	17	3	2
12	5	6	8	10	13	15	17	18	18	17	15	13	11	9	7	5
13	1	2	3	6	12	22	37	57	72	73	59	39	23	13	7	3
14			1	3	8	23	59	124	175	144	76	31	11	4	1	
15				1	1	2	4	7	12	18	28	39	47	50	45	35
16	1	1	2	3	5	8	12	18	25	32	37	38	35	28	21	15
17	1	2	3	6	12	20	33	49	61	63	54	38	24	14	8	4
18	1	2	5	9	18	33	54	74	81	69	47	28	15	8	4	2
19						1	2	3	7	13	24	41	61	75	73	57
20				1	1	2	4	8	15	26	43	61	72	67	51	33

*Trait levels are incremented by .4 units.

Table 6

Item Characteristics from BICAL Analysis

Item	Trait Level	Standard Error	Discrim.	Fit	Error Impact
1	1.71	.22	1.03	-0.85	.0
2	-1.19	.13	1.23	-2.35	.0
3	0.05	.15	1.14	-1.92	.0
4	-0.27	.14	0.21	7.55	.26
5	0.79	.17	1.04	-0.50	.0
6	0.66	.16	1.23	-2.10	.0
7	-0.01	.14	1.18	-1.76	.0
8	-1.27	.13	0.69	2.64	.07
9	0.46	.16	0.80	0.86	.04
10	1.57	.21	1.09	-0.77	.0
11	-1.15	.13	1.29	-2.91	.0
12	-1.35	.13	0.80	1.51	.04
13	-1.94	.13	1.27	-2.24	.0
14	-0.91	.13	1.48	-4.91	.0
15	1.41	.20	1.01	-1.15	.0
16	0.19	.15	1.03	-0.25	.0
17	-0.70	.13	1.21	-2.18	.0
18	-1.54	.13	1.30	-2.80	.0
19	2.16	.26	1.00	-1.10	.0
20	1.33	.20	1.09	-0.96	.0
Mean	0.00		1.03	-0.81	
S.D.	1.22		0.35	2.57	

Table 7

Item Characteristics from CREDIT Analysis

Item	Trait Level	Error	Fit to Scale
1	0.79	.08	-1.54
2	-0.92	.07	-4.99
3	0.03	.08	1.03
4	-0.14	.08	9.94
5	0.65	.09	2.49
6	0.43	.08	-2.39
7	0.21	.08	2.18
8	-0.86	.07	2.95
9	0.09	.08	2.45
10	1.23	.10	0.11
11	-0.90	.07	-7.14
12	-0.97	.07	-1.21
13	-1.22	.07	0.05
14	-0.68	.07	-6.76
15	0.90	.09	-0.26
16	0.15	.08	-0.46
17	-0.35	.07	3.02
18	-0.97	.07	-3.92
19	1.37	.10	-3.36
20	1.16	.10	-0.60
Mean	0.00	.08	-0.42
S.D.	0.83	.01	3.90

Table 8

Selection Rank of Items by
Respective Criteria

Item	Likert	Guttman	Logist	Bical	Credit	Rank Mean
1*	17.5	19	16	17	10	15.9
2	5	6	5	5	17	7.6
3	9.5	4	7	9	4	6.7
4*	20	11	20	20	20	20.0
5*	16	14.5	12	13	9	12.9
6*	11	14.5	14	12	8	11.9
7	7	2	8	8	7	6.4
8	17.5	9	10	4	3	8.7
9*	19	13	13	11	5	12.2
10*	13	18	18	16	13	15.6
11	3	5	2	6	19	7.0
12	14	10	9	3	2	7.6
13	4	8	4	1	1	3.6
14	1	1	1	19	18	8.0
15*	15	16.5	17	15	11	14.9
16*	9.5	12	11	10	6	9.7
17	7	3	6	7	14	7.4
18	2	7	3	2	16	6.0
19*	7	20	19	18	15	15.8
20*	12	16.5	15	14	12	13.9

*Response continuum reversed

Table 9

Rank Order Correlation Matrix
of Items Within Procedures

	GUTTMA.	LOGIST	BICAL	CREDIT
LIKERT	.576	.748	.356	-.304
GUTTMAN		.850	.497	-.066
LOGIST			.663	.002
BICAL				.430

Table 10

Characteristics of Method
Defined and Referent Scales

Scale	Mean	S.D.	Alpha
RANDOM	17.87	5.10	.861
BESTAVG	23.33	5.94	.866
WORSTAVG	16.35	4.74	.829
LIKERT	21.66	5.81	.868
NLIKERT	18.02	4.59	.790
GUT/LOGIST	23.33	5.94	.866
NGUT/NLOGIST	16.35	4.94	.829
BICAL	22.88	5.74	.849
NBICAL	16.80	4.73	.828
CREDIT	20.20	5.29	.824
NCREDIT	19.48	5.07	.842

Table 11

Correlation Matrix of Scores
for the Validation Sample

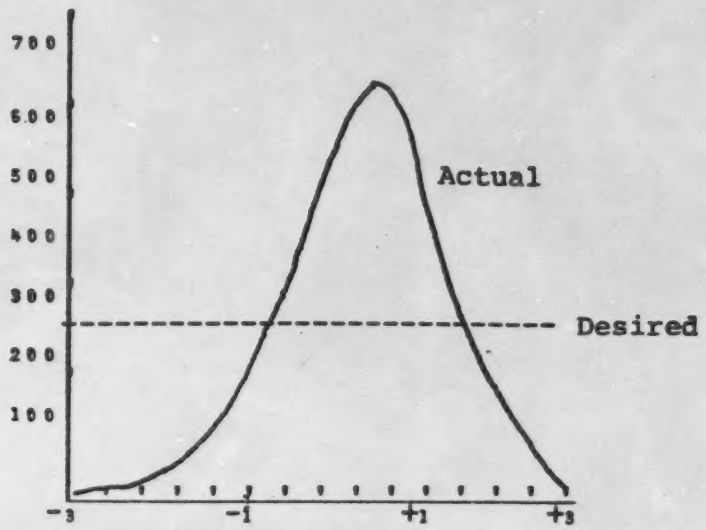
	LIKERT	GUT/LOG	BICAL	CREDIT
LIKERT	1.000	.968	.978	.884
NLIKERT	.767			
GUT/LOGIST		1.000	.980	.881
NGUT/NLOGIST		.675		
BICAL			1.000	.911
NBICAL			.745	
CREDIT				1.000
NCREDIT				.782
RANDOM	.895	.838	.871	.906
BESTAVG	.968	1.00	.980	.881
WORSTAVG	.755	.675	.725	.850

Table 12

Reliability and Validity
Characteristics of Scales

Scale	Alpha	Criterion Correl	Corrected for Reliabilities
RANDOM	.861	.665	.764
BESTAVG	.866	.648	.742
WORSTAVG	.829	.558	.653
LIKERT	.868	.699	.800
NLIKERT	.790	.532	.638
GUT/LOGIST	.866	.648	.742
NGUT/NLOGIST	.829	.558	.653
BICAL	.849	.646	.748
NBICAL	.828	.591	.692
CREDIT	.824	.572	.671
NCREDIT	.842	.686	.797

Figure 1: Comparison of Ideal and Actual Test Information Curves



REFERENCES

Andrich, D. Application of a psychometric rating model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 1978, 4, 581-594.

Andrich, D. Using latent trait measurement models to analyse attitudinal data: A synthesis of viewpoints. Paper presented at the 50th Anniversary of the Australian Council for Educational Research, Melborne, Australia, 1980.

Bejar, I.I. An application of the continuous response level model to personality measurement. Applied Psychological Measurement, 1977, 1, 509-521.

Benson, J., Crocker, L.M., & Ware, W.B. A comparison of three types of item analysis in test development using classical and latent trait methods. A paper presented at the Annual Meeting of the National Council on Measurement in Education, Toronto, 1978, (ERIC Document Reproduction Service No. ED163039).

Birnbaum, A. Statistical theory for logistic mental test models with a prior distribution of ability. Journal of Mathematical Psychology. 1969, 6, 258-276.

Bock, R.D. & Jones, L.V. The Measurement and Prediction of Judgment and Choice. San Francisco, Holden Day, 1968.

Dawes, R.M. Fundamentals of Attitude Measurement, New York: John Wiley, 1972.

Edwards, A.L. On Guttman's Scale analysis. Educational and Psychological Measurement. 1948, 8, 313-318.

Edwards, A.L. Techniques of Attitude Scale Construction. New York: Appleton-Century-Croft, Inc., 1957.

Edwards, A.L., & Kilpatrick, F.P.A. A technique for construction of attitude scales. Journal of Applied Psychology, 1948, 32, 374-384.

Goodenough, W.H.A. A technique for scale analysis. Educational and Psychological Measurement, 1944, 9, 139-150.

Gustafsson, J.-E. The Rasch model for dichotomous items: theory, applications, and a computer program. University of Goteborg, Institute of Education, 1977. (ERIC Document Reproduction Service No. ED154018).

Guttman, L.A. On Festinger's evaluation of scale analysis. Psychological Bulletin, 1947, 44, 451-465.

Hambleton, R.K. & Cook, L.L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hovland, C.I., & Sherif, M. Judgmental phenomena and scales of attitude measurement. Journal of Abnormal and Social Psychology, 1952, 47, 822-832.

Kolen, M.J. Comparison of Equipercentile, Linear, and Selected Latent Trait Methods for equating Forms and Levels of the Seventh Edition of the Iowa Tests of Educational Development. Doctoral dissertation, University of Iowa, 1979.

Liebermann, M. Estimation of moral judgment level using items whose alternatives form a graded scale. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1971. (ERIC Document Reproduction Service No. ED048341).

Lord, F.M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, N.J: Lawrence Erlbaum Associates, Publishers, 1980.

Masters, G.N. A Rasch model for partial credit scoring. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

Masters, G., Wright, B., & Ludlow, L. CREDIT: A Rasch program for ordered categories. University of Chicago, 1980.

Mehrens, W.A., & Lehmann, I.J. Measurement and Evaluation in Education and Psychology. 2nd ed. New York: Holt, Rinehart, and Winston, 1975.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K. & Bent, D.H. Statistical Package for the Social Sciences, 2nd ed. New York: McGraw-Hill Book Company, Inc., 1975.

Payne, D.A. The Assessment of Learning: Cognitive and Affective. Lexington: D.C. Heath & Company, 1974.

Russell, D., Peplau, L.A. & Cutrona, C.E. The Revised UCLA Loneliness Scale: concurrent and discriminant validity evidence. Journal of Personality and Social Psychology, 1980, 39, 472-280.

Russell, D., Peplau, L.A., & Ferguson, M.L. Developing a measure of loneliness. Journal of Personality Assessment, 1978, 42, 290-294.

Russell, D., Stephan, M., & Salih, F. Testing a cognitive model of loneliness. A paper presented at the Annual Meeting of the American Psychological Assocation, Los Angeles, 1981.

Ryan, J.P. & Hamm, D.W. Practical procedures for increasing reliability of classroom tests by using the Rasch model. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, 1976. (ERIC Document Reproduction No. ED129901).

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrics Monograph, 34, 1969.

Schmitt, N. Rasch analysis of the Central Life Interest Measure. Applied Psychological Measurement, 1981, 5, 3-10.

Schriesheim, C.A., & Hill, K.D. Controlling acquiesence response bias by item reversals: the effect on questionnaire validity. Educational and Psychological Measurement, 1981, 41, 1101-1114.

Strivastava, A.B.L., & Webster, H. An estimate of true score in the case of items scored on a continuous scale. Psychometrika, 1967, 32, 327-338.

Thorndike, R.L., & Hagen, E.P. Measurement and Evaluation in Psychology and Education, 4th ed., New York: John Wiley and Sons, 1977.

Thurstone, L.L., & Chave, E.J. The Measurement of Attitude. Chicago: University of Chicago Press, 1929.

Torgerson, W.S. Theory and Methods of Scaling. New York: John Wiley & Sons, Inc. 1958.

Wood, R.L., & Lord, F.M. A User's Guide to LOGIST. Research Memorandum RM-76-4. Princeton, N.J.: Educational Testing Service, 1976.

Wood, R.L., Wingersky, M.S. & Lord, F.M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (RM-76-6) Educational Testing Service, Princeton, New Jersey. June 1976, modified January 1978.

Wright, B.D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B.D., & Douglas, G.A. Best test design and self-tailored testing. Research Memorandum No. 19, Statistical Laboratory, Department of Education, University of Chicago, 1975a.

Wright B.D., & Douglas, G.A. Better procedures for sample-free item analysis. Research Memorandum No. 20, Statistical Laboratory, Department of Education, University of Chicago, 1975b.

Wright, B.D., & Douglas, G.A. Best procedures for sample-free item analysis. Applied Psychological Measurement, 1977a, 1, 281-294.

Wright, B.D., & Douglas, G.A. Conditional versus unconditional procedures for sample-free item analysis. Applied Psychological Measurement. 1977b, 37, 573-586.

Wright, B.D. & Mead, R.J. BICAL: Calibrating items and scales with the Rasch model. Research Memorandum No. 23A, University of Chicago, 1978.

Wright, B.D. & Stone, M.H. Best Test Design: Rasch Measurement. Chicago: Mesa Press, 1979.